An accessible, scalable ecosystem for enabling and sharing diverse mass spectrometry imaging analyses

Curt R. Fischer [b], Oliver Rübel [a], Benjamin P. Bowen [b,*]

[a] Computational Research Division, Lawrence Berkeley National Lab, USA
[b] Life Sciences Division, Lawrence Berkeley National Lab, One Cyclotron Road, Berkeley CA 94720, USA
[*] Corresponding author. E-mail address: BPBOWEN@LBL.GOV (B.P. Bowen).

January, 2016

Acknowledgment

This work was supported by and uses resources of Lawrence Berkeley National Lab's Laboratory Directed Research and Devel- opment (LDRD) support; the National Energy Research Scientific Computing Center (NERSC) supported by the Office of Science of the U.S. Department of Energy; and the Low-Dose Radiation Research of the Office of Science, Office of Biological and Environ- mental Research, of the U. S. Department of Energy under Contract No. DE-AC02-05CH11231. We thank S. Cholia and the Outreach, Software and Programming Group at NERSC for their ongoing ef- forts and support to help deliver scientific data and high- performance computing to science communities. We are thankful to Dr. Thomas Fehniger for his contribution of the histological im- age, and would like to dedicate this manuscript to his memory in recognition of his longstanding interest, collegiality, and support to the mass spectrometry imaging community.

Legal Disclaimer

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California.

An accessible, scalable ecosystem for enabling and sharing diverse mass spectrometry imaging analyses

Curt R. Fischer [b], Oliver Ruebel [a], Benjamin P. Bowen [b, *]

[a] Computational Research Division, Lawrence Berkeley National Lab, USA
[b] Life Sciences Division, Lawrence Berkeley National Lab, One Cyclotron Road, Berkeley CA 94720, USA
* Corresponding author.
E-mail address: BPBOWEN@LBL.GOV (B.P. Bowen).

Abstract

Mass spectrometry imaging (MSI) is used in an increasing number of biological applications. Typical MSI datasets contain unique, high-resolution mass spectra from tens of thousands of spatial locations, resulting in raw data sizes of tens of gigabytes per sample. In this paper, we review technical progress that is enabling new biological applications and that is driving an increase in the complexity and size of MSI data. Handling such data often requires specialized computational infrastructure, software, and expertise. OpenMSI, our recently described platform, makes it easy to explore and share MSI datasets via the web e even when larger than 50 GB. Here we describe the integration of OpenMSI with IPython notebooks for transparent, sharable, and replicable MSI research. An advantage of this approach is that users do not have to share raw data along with analyses; instead, data is retrieved via OpenMSI's web API. The IPython notebook interface provides a low-barrier entry point for data manipulation that is accessible for scientists without extensive computational training. Via these notebooks, analyses can be easily shared without requiring any data movement. We provide example notebooks for several common MSI analysis types including data normalization, plotting, clustering, and classification, and image registration.

Mass spectrometry is usually carried out on homogenized samples. Homogenization, whether achieved by mechanical pulverization or chemical extraction, destroys information on the spatial distribution of analytes in the sample. Mass spectrometry imaging (MSI) seeks to eliminate this information loss and to obtain both chemical and spatial information on analyzed samples. As such, it cannot rely on complete homogenization of samples.

Recent years have seen tremendous improvements in instrumental capabilities for MSI [22,47]. In this perspective, we discuss how the advances in instrument design and sample preparation have led to new challenges for MSI users, especially the size and complexity of the resulting datasets. We then show how OpenMSIea web-accessible repository for MSI data and a platform for sharing and analyzing dataein combination with sharable programming notebooks based on IPython can address these challenges, and demonstrate practical application of this idea by providing example notebooks performing two common types of MSI analyses. The article focuses primarily on soft-ionization techniques due to the greater chemical information content they provide.

1. Better instrumentation leads to bigger, more complex data

Here we review instrumental advances that are leading to increasing spatial resolution of MSI data and also the orthogonal instrumental improvements that increase chemical “resolution”. Higher spatial resolution leads to an increase in the number of pixels in an MSI image, and higher chemical resolution increases the complexity (or number) of the mass spectra recorded at each pixel. Simultaneously making use of improvements in both spatial and chemical resolution thus strongly increases the data size and complexity of MSI images, as we discuss below.

1.1. Instrumental improvements: spatial resolution

Spatial resolution in the single mm range, required for single-cell analysis of bacteria and also of many eukaryotic cells, has been generally inaccessible for MSI analyses that rely on soft ionization techniques such as matrix-assisted laser desorption (MALDI) or desorption via electrospray impact (DESI). Hard-ionization techniques, such as SIMS or nanoSIMS, easily reach these resolutions, and have been the subject of several recent reviews [26,41], but the harsh ionization conditions lead to analyte fragmentation into atoms or very small molecular fragments. Here our focus is on soft-ionization techniques, which can provide information on the abundance of intact metabolites, lipids, peptides, and proteins. Recent improvements discussed below are allowing even soft-ionization techniques to approach this single-cell limit. The x-axis of Fig. 1A shows obtainable spatial resolution for the most widely used ionization techniques.

1.1.1. Laser rastering techniques

The most popular and widespread means of collected spatially resolved mass spectra is by laser rastering over a sample surface. Matrix-assisted laser desorption ionization (MALDI) mass spectrometers usually use this imaging mode. Vendors usually fix the laser optics in a single position, and an XY-stage moves the sample across the incident laser beam. Spatial resolution can thus be limited by (a) the laser spot size and (b) the precision of the XY-stage [64]. Today's commercial MALDI-MS instruments have laser spot sizes that are in the range of 10e200 mm, and XY-stages with 1e10 mm translational precision.

Laser spot sizes thus usually constrain obtainable resolution. Oversampling techniques can improve spatial resolution to the limits of the XY-stage [23]. In Jurchen's implementation of this technique, a laser with a 100 x 200 mm spot size was repeatedly fired at a fixed position until the sample ions were no longer detected. A translation of the XY stage by 25 mm increments after analyte depletion brings fresh sample into the laser spot, allowing attribution of new signal to the 25 mm-region newly moved into the beam. Improved laser optics have also been reported that reduce laser spot sizes in MALDI-MS to about 2 mm [13,62]. A limitation of all rastering techniques is that increases in spatial resolution also increase data acquisition time.

“Ion microscopy”. Heeren and co-workers have accelerated MSI acquisition times by developing a unique imaging mode for MSI that relies on spatially resolved detection of ions ejected from a sample, rather than spatially resolved ionization. They term this mode “ion microscopy” and have demonstrated it using MALDI [22] ionization techniques. Their technique takes spatially-resolved mass spectra “inside of” the laser spot of a MALDI-type ionization system, allowing for much larger laser spot sizes (~200 mm) and faster acquisition times. Large sample images are constructed as mosaics of these single-spot images as the laser rasters across an image surface. Spatial resolution is thus independent of the laser spot size and is instead dictated by both the magnification inherent to the ion optics of the mass spectrometer, as well as the pixel sizes of the ion detector. Imaging resolution with ion microscopy of 6 mm has been reported. The same detector style can also be used for secondary-ion based ionization [28]; spatial resolving power of 7 mm was demonstrated.

1.1.2. Desorption electrospray ionization (DESI)

Desorption electrospray ionization simplifies sample prepration, and unlike many laser ionization techniques, can work at atmospheric pressure. In DESI, an electrospray of solvent droplets impacts the sample surface, causing desorption and ionization of analyte molecules. The electrospray is generated from two nested capillaries. The inner capillary contains the electrosprayed solvent, and the outer one contains a heated gas stream. The first example of using a solvent electrospray to perform MSI was reported in 2004 [51]. DESI does not require a matrix or initiator to absorb laser energy and initiate ionization.

Early examples of DESI MSI reported spatial resolution of 100e250 mm [52,58], but other studies have identified variables that control spatial resolution in DESI and improved resolution to 35 mm [8,10]. Variables strongly affecting spatial resolution include solvent flow rate, XY-stage step size, and the geometric orientation of the electrospray emitter relative to the mass spectrometer [8]. Additionally, the penetration of solvent from electrospray droplets into the sample can partially dissolve and distort the sample, decreasing spatial resolution, but altering solvent composition can ameliorate this effect. N,N-dimethylformamide:ethanol mixtures were superior to methanol:water mixtures in this regard. Use of

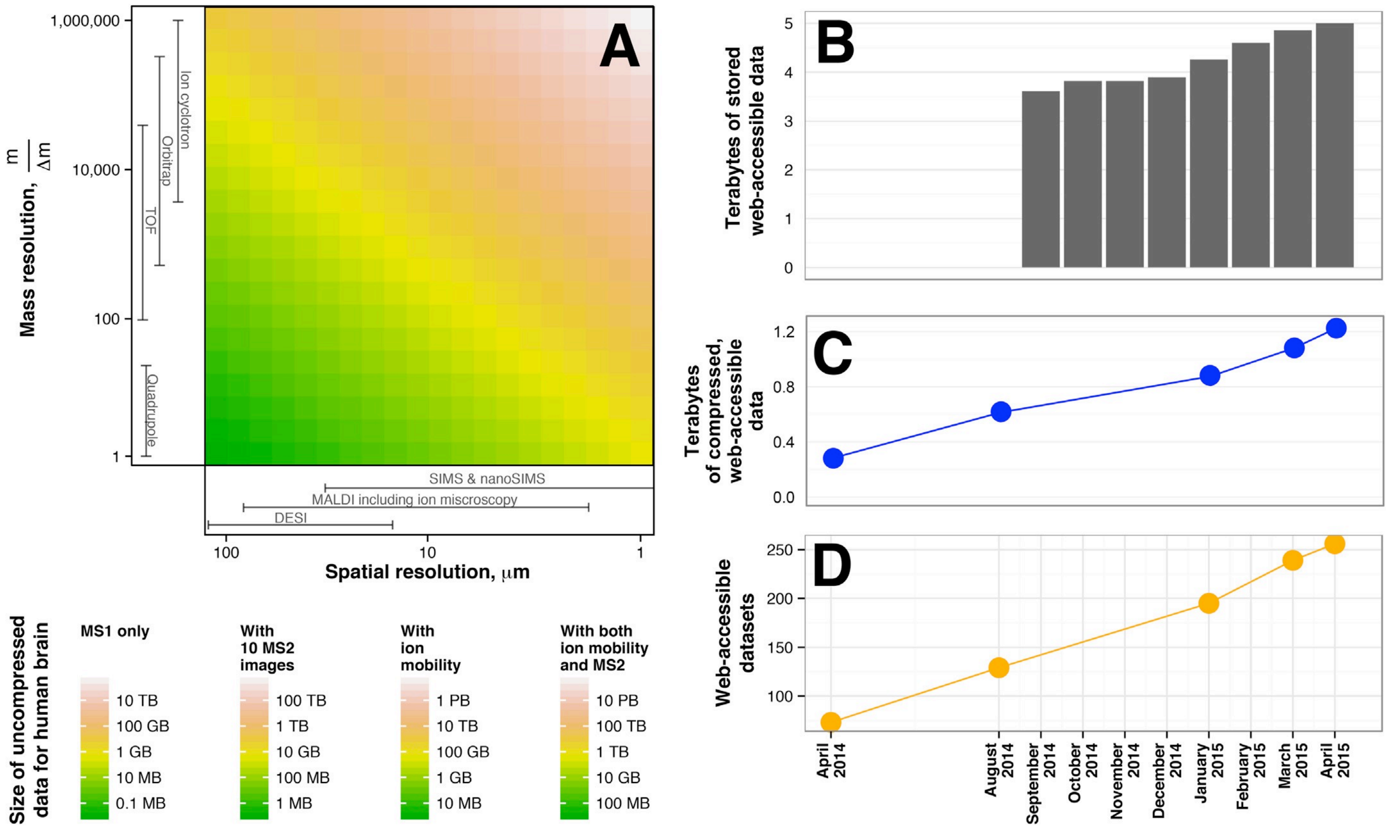

Fig. 1. Increased technological improvements are increasing the sizes for MSI datasets as illustrated by (A) an estimation of data size required for MSI analysis of human brain and as evidenced by (B) growth in real data stored in OpenMSI. Shown in (A) is a projected estimate for file size of an MSI data set of a single cross-section through a human brain using various techniques. As advances in spatial and mass resolution are combined with new dimensions including ion mobility separation and MS2, raw data sizes will increase dramatically. This growth is already evident today in the growth in the total size (B), terabytes of web-accessible data (C), and number of real MSI data sets (D) stored in OpenMSI.

"nanospray"-sized capillaries to direct the electrospray can improve resolution down to 12 mm [30].

1.1.3. Other ionization techniques

Although MALDI and DESI are the most widely used ionization techniques for mass spectrometry imaging, many groups continue to research other means for sample ionization and have reported impressive improvements. Several reviews have summarized progress in this area [12,15,16,45]. Most techniques rely on either laser shots, electrospray, or combinations of both to achieve ionization. Thus, obtainable spatial resolution is likely to be broadly similar to what has been achieved for MALDI or DESI.

1.2. Instrumental improvements in MSI: chemical resolution

The goal of MSI is the spatial mapping of the molecular composition of complex samples. Molecular composition must be inferred from mass spectra and knowledge of the bulk average composition of similar samples. We use the term "chemical "resolution" loosely to indicate the degree to which related compounds that originate from the same location of the sample can be distinguished.

1.2.1. High-resolution mass detectors

Mass spectrometry separates ions on the basis of their mass-to-charge ratio (m/z). In this context, resolution has a specific, precise meaning: how close can two ions be in their m/z and still be resolved by the detector? Widely available quadrupole detectors offer ~1 Da resolution, but time-of-flight (TOF) and Orbitrap detectors offer much improved mass resolution, in the 10e50 mDa range. FT-ICR detectors have demonstrated mass resolution in the single mDa range or lower. For the m/z ranges commonly observed in small-molecule mass spectrometry (ca. 30 Dae4000 Da), this means each spatial pixel contains ion intensities at 400,000 or more distinct m/z values. Obtainable m/z resolution for the mass analyzers commonly used in MSI is shown on the y-axis of Fig. 1A.

These high m/z resolving powers and high mass accuracies improve but still do not fully solve the task of identifying the chemical composition of a given ion [6,27]. Fortunately, other mass spectrometry tools that ameliorate this problem are increasingly being applied to MSI.

1.2.2. Tandem mass spectrometry and MS^n

Chemical information on the molecular structure of detected ions can also be obtained through fragmentation of detected ions, and observation of the mass spectrum of the resulting fragments. This process, termed tandem mass spectrometry or MS^2, is ubiquitous in the broader MS community but has seen less attention from imaging-focused researchers. Some instruments have the capability to fragment ions that are themselves fragments of an initially detected ion, and so on (MS^n) while at the same time offering high mass resolution. Instruments capable of MS^n analysis up to n ¼ 10 are commercially available. Some imaging applications of MS^2 have been reported [44], but in general this mode of mass spectrometry is not well-explored in MSI applications.

Depending on how data is acquired, MS^n adds dimensionality to the collected MSI images in several possible ways. Modern instruments can obtain MS^n data for fixed, predetermined precursor m/z values as well as for dynamic precursor m/z values determined during data acquisition via data-dependent fragmentation strategies. For example, instruments could be set to fragment the most abundant MS^1 ions in a given pixel, assuming that the ion has not already been fragmented in a nearby pixel. Some instruments allow further data-dependent dissection of MS^2 fragments into automatically chosen and acquired MS^3, …, MS^n spectra. These modes remain to be explored in MSI. Using such, data-depended data acquisition strategies results in the generation of varying numbers of MS^n spectra at possibly different precursor m/z values at each location. In contrast, when users define a predetermined list of MS^1 ions they wish to fragment, they can generate target lists of 1e~100 distinct m/z values that the instrument will fragment when detected. If these targeted fragmentations are run at every spatial pixel, there will be fixed set of multiple mass spectra for the same precursor m/z values at every pixel.

The limit on how many independent mass spectra can be acquired at each pixel is affected by several factors. In theory instrument software can limit the length of chosen "target lists", but practically in MSI, sample abundance and depth are likely limiting, as acquiring 100 spectra at each pixel would require (at least) 100 separate laser shots. The assumption that the data obtained from the nth shot remains representative of the sample surface as it existed at the first shot requires empirical validation.

1.2.3. Ion mobility separation (IMS)

A long-standing problem of mass spectrometry is that no matter how high an instrument's mass resolution, isomers cannot be resolved, because they have the same m/z. In non-imaging applications, isomers can be separated before mass spectral analysis, usually via liquid or gas chromatography. Chromatography is not possible in MSI, but ion mobility separation can resolve some isomeric species after ionization based on differential mobility of ions being accelerated through a dilute gas [24,35,49,52]. The rate of collisions between ions and gas molecules depends on molecular shape (specifically on collisional cross section), not only m/z. Thus, for every spectral scan at a particular pixel, IMS introduces a new dimension: drift time. In current commercially available setups, ions take milliseconds to travel the length of the drift tube. Since mass spectra can be acquired at a microsecond time scales, usually at least 200 spectra can be recorded as a function of drift time, at every pixel.

The next few years of MSI research are likely to see further advancements in chemical resolution ewhether IMS, MS^2 or higher-order fragmentation, or higher-resolution mass analyzers e combined with advancements in spatial resolution. The implications of these developments for the size of MSI datasets are shown in Fig. 1. The figure shows the raw (uncompressed, uncentroided) dataset size required to image a single 140 mm by 160 mm cross-section (or series of slices) of a human brain as a function of mass resolution and spatial resolution for several distinct scenarios. The assumed m/z acquisition range is from 50 Da to 4000 Da. The first colorbar shows data sizes for performing a single MS^1 scan at each pixel. At a spatial resolution of 10 mm and a mass resolution of 100,000, the size of the raw data can approach 1e10 terabytes. Adding 10 unique MS^2 acquisitions at each pixel, or 100 bins of time-resolved ion mobility, or both, at each pixel magnifies dataset sizes accordingly, with up to 1e10 petabytes required for such human brain-sized analyses.

The data size "disaster" can be ameliorated somewhat by smart representation of the obtained data. "Sparse" representations of the data e e.g., based on the detection and removal of data values describing background only, including recording mass spectra in centroid rather than profile mode e could slash the size of the dataset by 3e5 orders of magnitude. However, great care must be taken to avoid the loss of important information and current software from most major instrument vendors does not permit easy "sparsification" of MSI data except in the m/z dimension (i.e. centroiding). Sparsification in the spatial dimensions or along MS2 or ion mobility time dimensions are not supported in commonly used MSI file formats such as *.mzML, *.imzML, although work is ongoing in this area. However, even if sparse data representations reduce datasize by 10,000-fold, Fig. 1 implies that sparse data for a single slice of human brain could be up to 100 GB to 1 TB. Although mere storage of datasets of this size can be achieved relatively easily, sharing the data, as well as browsing, processing, and analyzing it, can be a considerable challenge.

2. Increasingly diverse applications demand diverse analyses

A second challenge arises from the different meanings that "analyzing" data takes in different applications of MSI [1]. The range of biological questions and problems to which MSI is being applied means that MSI users must be able to subject MSI data to increasingly diverse analysis types. In this section we profile several interesting, unique biological problems to which MSI has recently been successfully applied, with an emphasis on the data manipulations required for each problem.

2.1. Data reduction via centroiding and peak finding

A primary tool for reducing MSI dataset sizes is the centroiding of detected peaks in the m/z dimension. Modern instruments include algorithms for centroiding during data acquisition, but centroiding and peak grouping in MSI present a few special challenges. First, the m/z centroids found for the same ion can vary slightly from pixel to pixel due to instrumental noise and drift. These slightly different m/z values must be associated and corrected to a single m/z value for downstream statistical analyses to correctly use centroided peak values. This correction of m/z values is not well supported by current analysis tools and can lead to errors due to false matching of peaks from different ions that are close in m/z.

2.2. Physiological and anatomical mapping from spectral signatures

Mass spectra from distinct tissues types or organs can differ. Early MSI investigations manually identified particular ions whose distribution differed across organ types [25], but computational methods have the advantage of not relying on manual data inspection, identifying unexpected ions whose distribution corresponds to sample physiology. Unsupervised and supervised methods have both been used to identify distinct tissues or organs based on spectral signatures from the pixels of MSI images. Popular unsupervised tools for this analysis mode include principal components analysis (PCA), k-means clustering, non-negative matrix factorization (NMF), and maximum autocorrelation factor (MAF). Supervised tools widely used in MSI included linear discriminant analysis (LDA) or partial least squares analysis (PLS). Several more sophisticated techniques adapted to the specific structure of MSI data are also being developed [1,3].

A thorough and compelling recent example is provided by Hanrieder and coworkers' study of the effects of the cyanobacterial neurotoxin b-N-methylamino-L-alanine (BMAA) on the anatomy of rat brains by TOF-SIMS imaging. MAF was used for unsupervised classification of brain regions, and also for identification of ions whose spatial distribution and amount was changed by BMAA exposure relative to controls [14].

MSI can also reveal distinct physiological regions of a tissue that appear identical by traditional histological approaches. For example, a comparison of six different unsupervised methods for assessment of intra-tumor heterogeneity in human biopsies of myxofibrosarcomal lesions revealed that each method gave distinct results [19], but through "agreement analysis" intra-tumor heterogeneities that were robustly identified by five of the tested techniques could be identified.

2.3. Tracking the spatial distribution of target molecules

Another important analysis mode relies on external identification of physiological regions of interest, followed by identifying the distribution of specific molecules detected by MSI in these regions. An early example was Khatib-Shahidi's study on the whole-body distribution of the antipsychotic benzodiazapene drug olanzapine and its metabolites in whole rat sagittal tissue sections. MSI clearly showed concentration of two drug metabolites in the bladder and liver, while the unmetabolized drug was more broadly distributed across the body [25]. MSI also showed that an inhaled dose of the bronchodilator ipratropium bromide concentrated in regions of high inflammation and cell density and away from epithelial tissue in airway biopsies from human patients [11].

This mode of analysis was also used to localize the tissue types and regions where biosynthesis of naturally produced drug podophyllotoxin [33]. The drug was known to be naturally produced by several species of the Podophyllum genus of eudicot herbs. MSI showed the drug precursor magnoflorine was concentrated in the epidermal tissue and emerging root tips of P. hexandrum rhizomes, but in different regions of P. peltatum rhizomes. These differing distributions across species coincided with the tissue distribution of mRNAs for two cytochrome P450 enzymes as determined by microdissection and RT-PCR, thus supporting a role for these P450s in the biosynthesis of magnoflorine and thus of podophyllotoxin [33].

Other recent MSI investigations have revealed differences in the sub-dermal profiles of the topical anaesthic lidocaine and its metabolites in pig ears [9], as well as investigating the metabolic interactions in live, mixed-species microbial colonies growing on agar plates [57] at resolutions of ~100 mm.

These modes of analysis all used pre-existing knowledge on drug metabolism or biosynthesis to target m/z values of interest. Tools for automatic identification of ions whose spatial distribution correlates with target ions (e.g. LDA or PLS in the m/z dimension rather than the spatial dimension) could help uncover unknown metabolites derived from or leading to important drugs.

2.4. Quantifying the rates of protein and metabolite turnover

MSI can be coupled with stable isotope labeling strategies to study protein and metabolite turnover in tissue or biofilm. One recent study applied this technique to study differences in phospholipid turnover in brain tissue sections from mice into which tumor lesions were surgically implanted. Five days after switching the mice's water supply to an 8% D_2O labeled feed, MSI analysis revealed that phospholipids containing saturated and monounsaturated fatty acids were synthesized at faster rates in tumor regions than in non-tumor regions. K-means clustering also revealed variations in lipid production among several different types of non-tumor regions in the brain [31]. Isotope labeling coupled to MSI via nanoSIMS also localized protein turnover to the tips of intracellular stereocilial fibers of inner ear cells of adult mice [63].

Key requirements for analysis of isotope labeling data in MSI include the ability to automatically identify peaks in the m/z dimension that are related by isotopic substitution. Such algorithms are routinely applied to LC-MS data, but the lack of chromatographic separation in MSI often means that the isotopic envelope of a given molecular species will be overlapping with those for several other molecular species. The future is likely to bring different, even more complex analysis types.

3. OpenMSI þ IPython notebooks enable sharable, collaborative data analysis on large scale MSI data sets

Thus far in this article we have argued that (i) advancing technology is rapidly increasing the size of MSI datasets (Fig. 1a), and (ii) the application of MSI to diverse problems in the biological sciences necessitates a capability for diverse types of data analysis. OpenMSI [46] is a web-based repository of MSI datasets as well as a platform for data analysis and sharing that addresses both of these challenges. OpenMSI is publically accessible at https://openmsi.nersc.gov and powered by supercomputing infrastructure at the National Energy Research Supercomputing Center (NERSC). NERSC runs several world-class supercomputing systems, all of which are available to OpenMSI users. An example is the Edison XC30 computer, which features 133,824 compute cores, 357 terabytes of memory, and 7.56 petabytes of online disk storage with a peak I/O bandwidth of 168 gigabytes (GB) per second. Edison has a theoretical peak performance of 2.57 petaflops/second.

Since its release in 2013, OpenMSI usage has increased steadily (Fig. 1bed). Over 5 TB of MSI data (a combination of both archived raw as well as converted files using OpenMSI's data in HDF5 format) are now stored in OpenMSI. Already more than 1.3 TB of MSI and analysis data, stored in more than 250 HDF5 files using lossless data compression, are accessible via the web through (i.e., the uncompressed, raw data volume is on the order of ~3.5 TB) (Fig. 1c,d).

In addition to the integrated web-based data sharing, processing, and visualization capabilities, OpenMSI provides an easy-to-use, powerful Web API that enables users to programmatically access data via OpenMSI remotely. OpenMSI's Web API consists of just five simple functions, i) qmetadata to retrieve metadata only, ii) qmz to retrieve information about data axes, and iii e v) qslice, qspectrum, and qcube which provide easy-to-use support for the three most common selective read patterns, i.e., read ion image slices, read m/z spectra and read arbitrary subcubes of the data. Together, these functions provide full access to the data, including metadata and raw MSI and derived analysis data. The basic methods are simple and can be effectively encoded in URL patterns. For further details see [46].

Using the OpenMSI Web API users can access MSI data remotely, e.g., via IPython notebooks. IPython notebook is a web-browser based interface to a Python interpreter that facilitates interactive coding and code sharing [43]. IPython notebooks combine code execution, rich text, mathematics, plots and rich media in a single environment and provide an agile tool for exploratory computation and data analysis. Programmable notebooks make scientific analysis easily reproducible by combining the state of the analysis, code, and documentation of the analysis steps in a single, human-readable document. This interface lowers the barriers for effective data manipulation in Python, permitting users with only a basic familiarity with Python to perform advanced data analyses. Python is a high-level interpreted language with a growing set of publicly available tools for common manipulations of multivariate and imaging data such as SciPy [18], scikit-learn [42], and scikit-image [53].

IPython notebooks can be easily shared and edited via version-control systems such as Git (www.github.com). Using the advanced collaboration features of GitHub and other online code repositories, users can collaborative develop and share analytics, adapt analytics for new applications, and record the provenance of the development and changes of analyses. Users without any programming skills can easily view analysis notebooks online using only a web browser via the IPython tool nbviewer (nbviewer.ipython.org).

The combination of OpenMSI, IPython notebooks, and online source code repositories achieves cross-platform, scalable access to large MSI datasets, version-controlled sharing of analysis methods, and easy public viewing and options (Fig. 2). This workflow enables the collaborative development of sophisticated workflows for analyzing MSI data, whether the end application be in research or clinical environments [56].

We here demonstrate the application of this approach to the visualization and analysis of public OpenMSI data by clustering- and matrix-factorization-based techniques. We have shared the analysis notebooks, which include the analysis code and detailed documentation, with the public via GitHub. No additional data is required for interested readers to execute the notebooks as all data is retrieved at runtime via OpenMSI. While we focus in this article on the use of Python, the proposed approach also extends directly to other programming languages, e.g., R and Julia. The recently released Jupyter notebook and multi-user server JupyterHub (https://jupyter.org/) have evolved from IPython and provide a language-agnostic environment for development of sharable, programmable science notebooks.

3.1. Loading and viewing OpenMSI data in IPython

A key advantage of using OpenMSI for MSI data analysis is i) that only as much data as is required for the desired analysis needs to be loaded to IPython, ii) selected data subsets can be retrieved fast, and iii) users do not need to share data files along with analysis notebooks, but the data is retrieved on request via OpenMSI. For plotting of chosen ion images and mass spectra at chosen pixels, the full data set remains stored and accessible at the OpenMSI server, and only chosen data subsets are passed to IPython. The capabilities of IPython make it easy to generate complex, publication-quality data visualizations. An IPython notebook at http://tinyurl.com/openmsi-nb1 shows how to load, plot, and zoom in on images and spectra from a publicly available OpenMSI dataset, in this case for a NIMS image of a mouse brain cross-section acquired using a TOF detector (http://tinyurl.com/openMSI-brain). The plots prepared entirely programmatically via this notebook are shown in Fig. 3 and Fig. 4.

3.2. K-means clustering on OpenMSI data

To provide an example of a more complex analysis workflow using IPython notebooks and OpenMSI, we provide a second notebook at http://tinyurl.com/openmsi-nb2 that uses the same publicly available OpenMSI dataset to perform filtering to eliminate pixels where mass spectra arise predominantly from background (matrix) ions or from combinations of background and sample ions. Then, we apply non-negative matrix factorization [19] to the filtered data set to further identify pixels in the "clean" image that are related to each other by spectral similarity. An overview of a set of components of the NMF matrix factorization is shown in Fig. 4. A similar analysis was recently presented by Yang [61].

3.3. Simple image registration of MSI and optical images

The task of finding the best way to "line up" the multiple images of the same object is known as "image registration", and can be difficult for tasks where the intensity, contrast, and resolution all vary strongly between the different images. This problem is referred to as multimodal image registration, and is a problem for comparing ion images from an MSI experiment to other image sources, such as optical microscopy. In a third example at http://tinyurl.com/openmsi-nb3, we align an optical image with an MSI image. We use the publically available rat lung dataset from OpenMSI and an associated image of the same sample obtained by optical microscopy. This image is an "H&E" image, i.e. a histologically fixed sample stained with hematoxylin and eosin and was kindly provided by Dr. Thomas Fehniger (Center of Excellence in Biological and Medical Mass Spectrometry, Lund, Sweden) [54]. For users interested in replicating the code, the image will be provided by request.

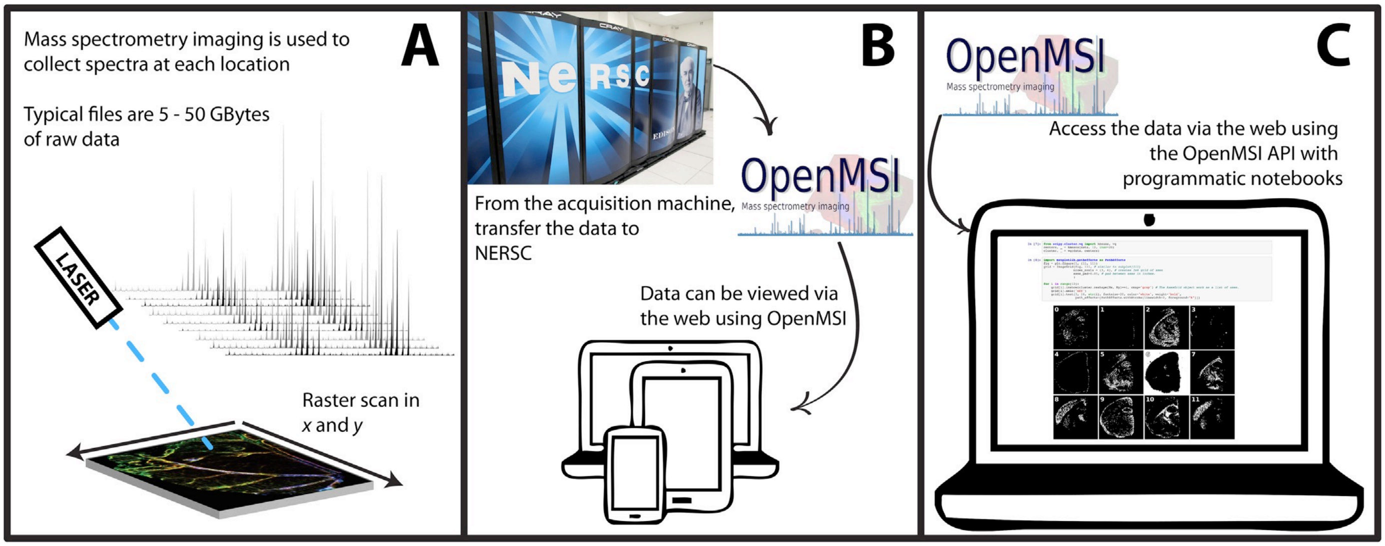
A
Mass spectrometry imaging is used to collect spectra at each location
Typical files are 5 - 50 GBytes of raw data
LASER
Raster scan in *x* and *y*
B
NERSC
OpenMSI
Mass spectrometry imaging
From the acquisition machine, transfer the data to NERSC
Data can be viewed via the web using OpenMSI
C
OpenMSI
Mass spectrometry imaging
Access the data via the web using the OpenMSI API with programmatic notebooks
0
1
2
3
4
5
6
7
8
9
10
11

Fig. 2. Illustration of the workflow for customizable mass spectrometry imaging analysis using OpenMSI and sharable IPython analysis notebooks. Panel A represents mass spectrometry image acquisition; data is generated that is often tens of GB or more per file. This data is transferred to OpenMSI, visualized with web browser based tools, and shared with users of OpenMSI (Panel B). IPython notebooks and the OpenMSI web API enable a scientist to remotely perform advanced programmable analytics (Panel C). Notebooks are easily sharable via public version-control tools such as GitHub and can be developed collaboratively. Since data is accessed on-demand via OpenMSI's API, sharing notebooks does not require copying or duplicating data.

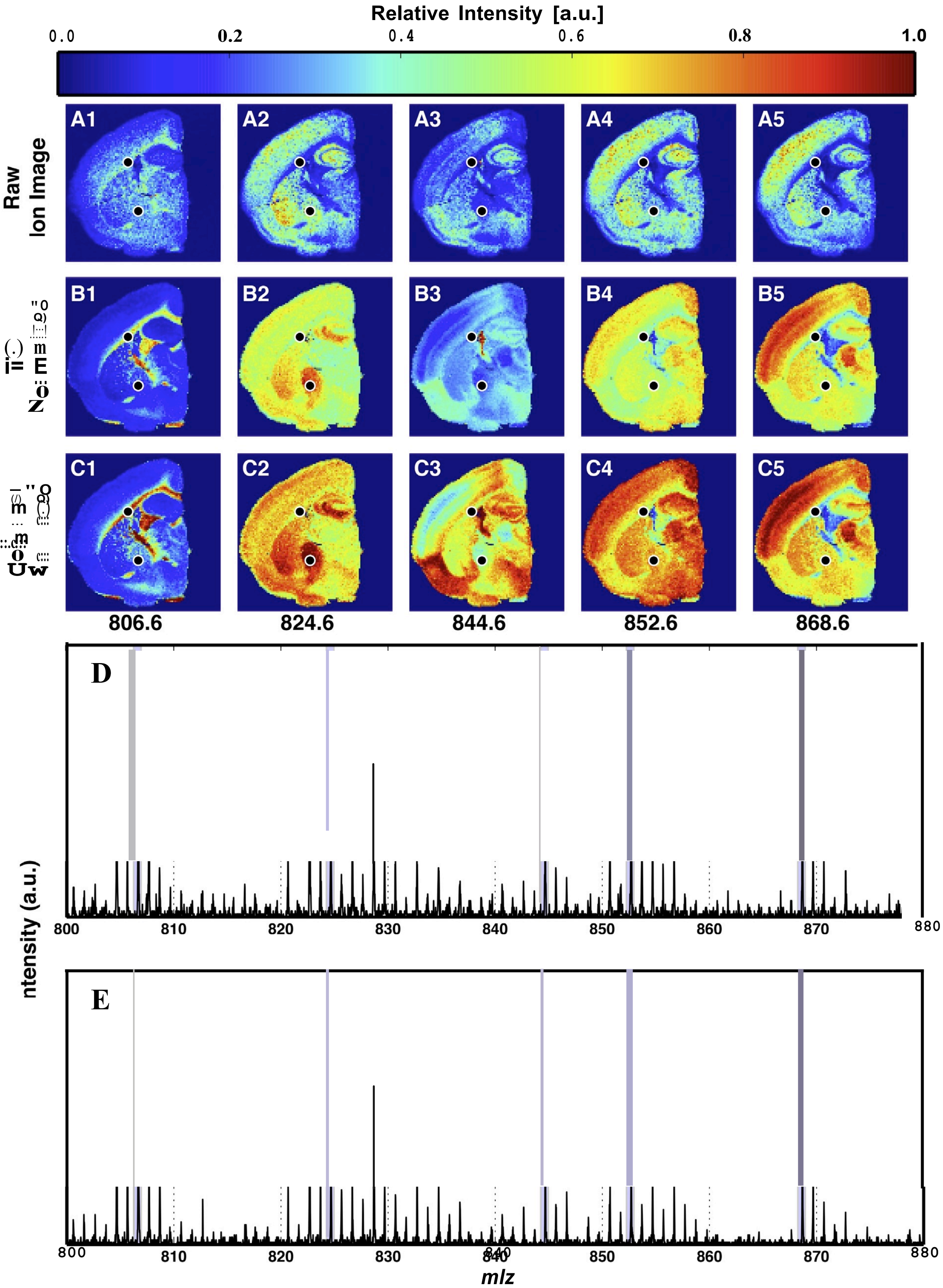

Fig. 3. Ion images and spectra created programmatically from a publicly available OpenMSI dataset using the !Python notebook at http:jjtinyurl.comjopenmsi-nb1. (A) Ion images. The raw ion images are shown in row A. In row B the same ions are normalized by the total intensity of just the chosen ions. Row Cis a contrast-enhanced version of Row Bin which high intensities are compressed, allowing the colormap to show more variation among low-intensity regions. Black dots on each image can annotate pixel locations of interest. (D) and (E) Mass spectra for the locations marked with black dots in (A). The *mjz* regions integrated to form the ion images in (A) are highlighted in gray. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

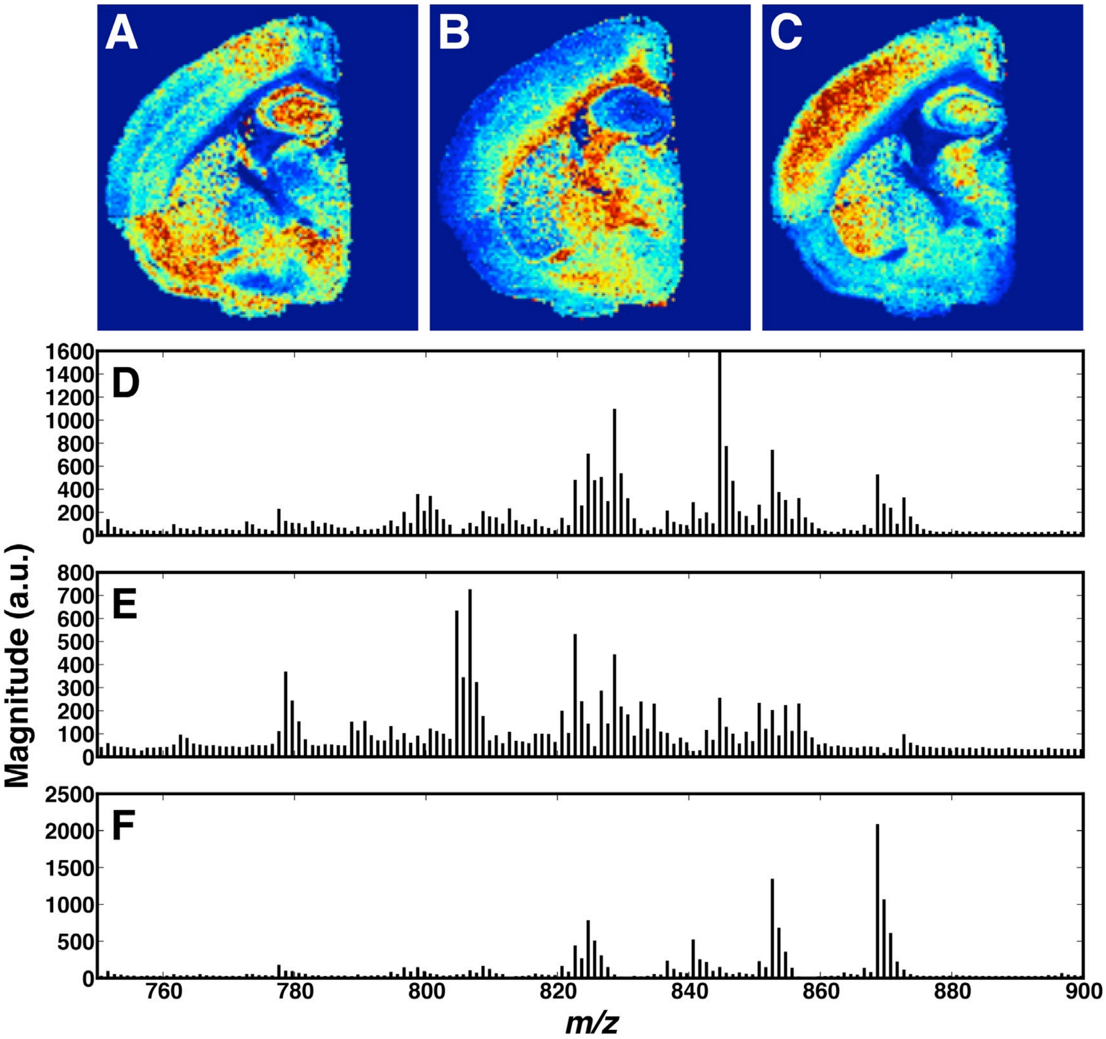

Fig. 4. Programmatic visualization illustrating the use of dimensionality reduction using non-negative matrix factorization (NMF) on publicly available OpenMSI data using an IPython notebook (http://tinyurl.com/openmsi-nb2). (A, B, and C): visualization of the first three NMF components showing the spatial component coefficients. (D, E, and F): Spectrum of ion-component coefficients (loadings) of m/z values for the first three NMF components, following [61].

4. Conclusions

In this article we have demonstrated that improving instrumental technology is increasing the size of MSI datasets. This trend is driven by increasing spatial resolution accessible via MALDI, ion microscopy, DESI, and other techniques, and also by increasing chemical resolution via the use of higher-resolution mass analyzers, as well as increased use of MS2 and ion mobility separation. We have also shown that MSI is being applied to an increasingly diverse set of problems in the biological sciences. This in turn requires a capability for easy performance of diverse types of analysis on MSI data. OpenMSI can store large MSI datasets and allow easy browsing and inspection via the web. When used in conjunction with IPython notebooks, all of the advanced data processing capabilities of Python, SciPy, and scikit-image, scikit-learn, and scikit-statmodels and many other available analysis packages can be easily applied to MSI data. OpenMSI allows easy sharing of MSI via the web. When coupled with the IPython notebook concept, complex analyses on MSI data can also be easily shared, reviewed, and collaboratively develop by multiple investigators, labs, and all members of the growing MSI community. While we focused in this article on IPython notebooks, the recent development of Jupyter allows the direct extension of the proposed approach to other programming languages, such as R and Julia. For these reasons, we hope these and similar tools (e.g., http://www.maldi-msi.org/) will be increasingly used by scientists interested in MSI.

Acknowledgments

This work was supported by and uses resources of Lawrence Berkeley National Lab's Laboratory Directed Research and Development (LDRD) support; the National Energy Research Scientific Computing Center (NERSC) supported by the Office of Science of the U.S. Department of Energy; and the Low-Dose Radiation Research of the Office of Science, Office of Biological and Environmental Research, of the U. S. Department of Energy under Contract No. DE-AC02-05CH11231. We thank S. Cholia and the Outreach, Software and Programming Group at NERSC for their ongoing efforts and support to help deliver scientific data and high-performance computing to science communities. We are thankful to Dr. Thomas Fehniger for his contribution of the histological image, and would like to dedicate this manuscript to his memory in recognition of his longstanding interest, collegiality, and support to the mass spectrometry imaging community.

References

[1] T. Alexandrov, Maldi imaging mass spectrometry: statistical data analysis and current computational challenges, BMC Bioinforma. 13 (Suppl. 16) (2012) S11.

[3] T. Alexandrov, J.H. Kobarg, Efficient spatial segmentation of large imaging mass spectrometry datasets with spatially aware clustering, Bioinformatics 27 (13) (2011) i230ei238.

[6] B.P. Bowen, C.R. Fischer, R. Baran, J.F. Banfield, T. Northen, Improved genome annotation through untargeted detection of pathway-specific metabolites, BMC Genom. 12 (Suppl. 1) (2011) S6.

[8] D.I. Campbell, C.R. Ferreira, L.S. Eberlin, R.G. Cooks, Improved spatial resolution in the imaging of biological tissue using desorption electrospray ionization, Anal. Bioanal. Chem. 404 (2) (2012) 389e398.

[9] J. D'Alvise, R. Mortensen, S.H. Hansen, C. Janfelt, Detection of follicular transport of lidocaine and metabolism in adipose tissue in pig ear skin by DESI mass spectrometry imaging, Anal. Bioanal. Chem. 406 (15) (2014) 3735e3742.

[10] L.S. Eberlin, C.R. Ferreira, A.L. Dill, D.R. Ifa, L. Cheng, R.G. Cooks, Nondestructive, histologically compatible tissue imaging by desorption electrospray ionization mass spectrometry, ChemBioChem 12 (14) (2011) 2129e2132.

[11] T.E. Fehniger, A. Vegvari, M. Rezeli, K. Prikk, P. Ross, M. Dahlbäck, G. Edula, R. Sepper, G. Marko-Varga, Direct demonstration of tissue uptake of an inhaled drug: proof-of-principle study using matrix-assisted laser desorption ionization mass spectrometry imaging, Anal. Chem. 83 (21) (2011) 8329e8336.

[12] T. Greer, R. Sturm, L. Li, Mass spectrometry imaging for drugs and metabolites, J. Proteom. 74 (12) (2011) 2617e2631.

[13] S. Guenther, M. Koestler, O. Schulz, B. Spengler, Laser spot size and laser power dependence of ion formation in high resolution MALDI imaging, Int. J. Mass Spectrom. 294 (1) (2010) 7e15.

[14] J. Hanrieder, L. Gerber, A. Persson Sandelius, E.B. Brittebo, A.G. Ewing, O. Karlsson, High resolution metabolite imaging in the hippocampus following neonatal exposure to the environmental toxin BMAA using ToF-SIMS, ACS Chem. Neurosci. 5 (7) (2014) 568e575.

[15] G.A. Harris, L. Nyadong, F.M. Fernandez, Recent developments in ambient ionization techniques for analytical mass spectrometry, Analyst 133 (10) (2008) 1297e1301.

[16] C. Ibanez, V. Garcıa-Canas, A. Valdes, C. Simo, Direct mass spectrometry-based approaches in metabolomics, Fundam. Adv. Omics Technol. Genes Metabolites 63 (2014) 235.

[18] E. Jones, T. Oliphant, P. Peterson, et al., SciPy: Open Source Scientific Tools for Python, 2001.

[19] E.A. Jones, A. van Remoortere, R.J. van Zeijl, P.C. Hogendoorn, J.V. Bovee, A.M. Deelder, L.A. McDonnell, Multiple statistical analysis techniques corroborate intratumor heterogeneity in imaging mass spectrometry datasets of myxofibrosarcoma, PLoS One 6 (9) (2011) e24913.

[22] J.H. Jungmann, D.F. Smith, L. MacAleese, I. Klinkert, J. Visser, R.M. Heeren, Biological tissue imaging with a position and time sensitive pixelated detector, J. Am. Soc. Mass Spectrom. 23 (10) (2012) 1679e1688.

[23] J.C. Jurchen, S.S. Rubakhin, J.V. Sweedler, Maldi-MS imaging of features smaller than the size of the laser beam, J. Am. Soc. Mass Spectrom. 16 (10) (2005) 1654e1659.

[24] H. Kettling, S. Vens-Cappell, J. Soltwisch, A. Pirkl, J. Haier, J. Müthing, K. Dreisewerd, Maldi mass spectrometry imaging of bioactive lipids in mouse brain with a synapt g2-s mass spectrometer operated at elevated pressure: improving the analytical sensitivity and the lateral resolution to ten micrometers, Anal. Chem. 86 (15) (2014) 7798e7805.

[25] S. Khatib-Shahidi, M. Andersson, J.L. Herman, T.A. Gillespie, R.M. Caprioli, Direct molecular analysis of whole-body animal tissue sections by imaging Maldi mass spectrometry, Anal. Chem. 78 (18) (2006) 6448e6456.

[26] M.R. Kilburn, P.L. Clode, Elemental and isotopic imaging of biological samples using nanosims, in: Electron Microscopy, Springer, 2014, pp. 733e755.

[27] T. Kind, O. Fiehn, Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry, BMC Bioinforma. 8 (1) (2007) 105.

[28] A. Kiss, J.H. Jungmann, D.F. Smith, R.M. Heeren, Microscope mode secondary ion mass spectrometry imaging with a Timepix detector, Rev. Sci. Instrum. 84 (1) (2013) 013704.

[30] J. Laskin, B.S. Heath, P.J. Roach, L. Cazares, O.J. Semmes, Tissue imaging using nanospray desorption electrospray ionization mass spectrometry, Anal. Chem. 84 (1) (2011) 141e148.

[31] K.B. Louie, B.P. Bowen, S. McAlhany, Y. Huang, J.C. Price, J.-h. Mao, M. Hellerstein, T.R. Northen, Mass spectrometry imaging for in situ kinetic histochemistry, Sci. Rep. 3 (2013).

[33] J.V. Marques, D.S. Dalisay, H. Yang, C. Lee, L.B. Davin, N.G. Lewis, A multiomics strategy resolves the elusive nature of alkaloids in podophyllum species, Mol. Biosyst. 10 (11) (2014) 2838e2849.

[35] J.A. McLean, W.B. Ridenour, R.M. Caprioli, Profiling and imaging of tissues by imaging ion mobility-mass spectrometry, J. Mass Spectrom. 42 (8) (2007) 1099e1105.

[41] V. Orphan, C. House, Geobiological investigations using secondary ion mass spectrometry: microanalysis of extant and paleo-microbial processes, Geobiology 7 (3) (2009) 360e372.

[42] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825e2830.

[43] F. Perez, B.E. Granger, IPython: a system for interactive scientific computing, Comput. Sci. Eng. 9 (3) (May 2007) 21e29.

[44] D.A. Pirman, R.F. Reich, A. Kiss, R.M. Heeren, R.A. Yost, Quantitative Maldi tandem mass spectrometric imaging of cocaine from brain tissue with a deuterated internal standard, Anal. Chem. 85 (2) (2012) 1081e1089.

[45] B. Prideaux, M. Stoeckli, Mass spectrometry imaging for drug distribution studies, J. proteom. 75 (16) (2012) 4999e5013.

[46] O. Rübel, A. Greiner, S. Cholia, K. Louie, E.W. Bethel, T.R. Northen, B.P. Bowen, Openmsi: a high-performance web-based platform for mass spectrometry imaging, Anal. Chem. 85 (21) (2013) 10354e10361.

[47] K. Schwamborn, R.M. Caprioli, Molecular imaging by mass spectrometrydlooking beyond classical histology, Nat. Rev. Cancer 10 (9) (2010) 639e646.

[49] J. Stauber, L. MacAleese, J. Franck, E. Claude, M. Snel, B.K. Kaletas, I.M. Wiel, M. Wisztorski, I. Fournier, R.M. Heeren, On-tissue protein identification and imaging by Maldi-ion mobility mass spectrometry, J. Am. Soc. Mass Spectrom. 21 (3) (2010) 338e347.

[51] Z. Takats, J.M. Wiseman, R.G. Cooks, Ambient mass spectrometry using desorption electrospray ionization (desi): instrumentation, mechanisms and applications in forensics, chemistry, and biology, J. Mass Spectrom. 40 (10) (2005) 1261e1275.

[52] P.J. Trim, C.M. Henson, J.L. Avery, A. McEwen, M.F. Snel, E. Claude, P.S. Marshall, A. West, A.P. Princivalle, M.R. Clench, Matrix-assisted laser desorption/ionization-ion mobility separation-mass spectrometry imaging of vinblastine in whole body tissue sections, Anal. Chem. 80 (22) (2008) 8628e8634.

[53] S. Van Der Walt, J.L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J.D. Warner, N. Yager, E. Gouillart, T. Yu, Scikit-image: image processing in python, PeerJ 2 (2014) e453.

[54] A.. Vegvari, T.E. Fehniger, L. Gustavsson, A. Nilsson, P.E. Andren, K. Kenne, J. Nilsson, T. Laurell, G. Marko-Varga, Essential tactics of tissue preparation and matrix nano-spotting for successful compound imaging mass spectrometry, J. Proteomics 73 (6) (2010) 1270e1278.

[56] K.A. Veselkov, R. Mirnezami, N. Strittmatter, R.D. Goldin, J. Kinross, A.V. Speller, T. Abramov, E.A. Jones, A. Darzi, E. Holmes, et al., Chemo-informatic strategy for imaging mass spectrometry-based hyperspectral profiling of lipid signatures in colorectal cancer, Proc. Natl. Acad. Sci. 111 (3) (2014) 1216e1221.

[57] J. Watrous, P. Roach, B. Heath, T. Alexandrov, J. Laskin, P.C. Dorrestein, Metabolic profiling directly from the petri dish using nanospray desorption electrospray ionization imaging mass spectrometry, Anal. Chem. 85 (21) (2013) 10385e10391.

[58] J.M. Wiseman, D.R. Ifa, Q. Song, R.G. Cooks, Tissue imaging at atmospheric pressure using desorption electrospray ionization (DESI) mass spectrometry, Angew. Chem. Int. Ed. 45 (43) (2006) 7188e7192.

[61] Jiyan Yang, Oliver Rubel, Prabhat, Michael W. Mahoney, Ben P. Bowen, Identifying important ions and positions in mass spectrometry imaging data using CUR matrix decompositions, Anal. Chem. 87 (9) (2015) 4658e4666.

[62] A. Zavalin, J. Yang, K. Hayden, M. Vestal, R.M. Caprioli, Tissue protein imaging at 1 mm laser spot diameter for high spatial resolution and high imaging speed using transmission geometry Maldi ToF MS, Anal. Bioanal. Chem. 407 (8) (2015) 2337e2342.

[63] D.-S. Zhang, V. Piazza, B.J. Perrin, A.K. Rzadzinska, J.C. Poczatek, M. Wang, H.M. Prosser, J.M. Ervasti, D.P. Corey, C.P. Lechene, Multi-isotope imaging mass spectrometry reveals slow protein turnover in hair-cell stereocilia, Nature 481 (7382) (2012) 520e524.

[64] Caprioli, Richard M. Farmer, Terry B. Gile, Jocelyn, Molecular imaging of biological samples: localization of peptides and proteins using MALDI-TOF MS, Anal. Chem. 69 (23) (1997) 4751e4760.